DOCUMENT RESUME

ED 159 196 TM 007 438

AUTHOR Pine, Steven M.
TITLE Applications of Item Response Theory to the Problem oF Test Bias.
SPONS AGENCY Office of Naval Research, Washington, D.C. Personnel and Training Branch.
BUREAU NO NR 150-383
PUB DATE [76]
CONTRACT N00014-76-C-0244
NOTE 11p.

EDRS PRICE MF-$0.83 HC-$1.67 Plus Postage.
DESCRIPTORS Complexity Level; *Culture Free Tests; Factor Analysis; *Item Analysis; Racial Differences; Secondary Education; *Test Bias; Test Construction; *Test Items; *Theories; Verbal Tests
IDENTIFIERS *Classical Test Theory; *Item Characteristic Curve Theory; Latent Trait Theory

ABSTRACT

Latent trait theory is used as a basis for a definition of item bias, and an experiment is described in which a test of vocabulary items was not found to be biased when administered to two groups of high school students, one group of 58 black students and the other 168 white students. The experiment for the detection of bias began with separate factor analyses of 45 verbal items for each of the two groups of students. Both analyses were judged to show that the set of items was unidimensional. Then scatter plots were made comparing item difficulty parameters for the two groups. The scatter plot was judged to be linear which was taken to mean that none of the items were biased. Examples of these items are given. The author believes that classical test theory is inadequate for the detection of bias in test items and prefers the three-parameter item characteristic curve model for that puppose. (CTM)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U S DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

ED159196

TM007 438

Applications of Item Response Theory to the Problem of Test Bias

by
Steven M. Pine
University of Minnesota

Introduction

One of the most challenging and important issues facing test developers and users today is whether or not tests are biased against certain minority groups, and if so, what to do about it. In recent years, there has been a plethora of research activity to study and reduce bias and unfairness in various settings. For the most part, these efforts have been dismal failures. A careful review of research in this area indicates that one possible reason for this lack of progress is the fact that almost all the research on test bias and fairness has been carried out within the context of classical test theory.

In his recent review of test theory, Lumsden (1976) refers to the true score model of classical test theory as the "Model-T Theory" and suggests that classical test theory has failed as a theory. Therefore, it makes little sense to continue working within this antiquated framework to resolve problems as complex as those involved in the bias issue. Not only is classical test theory antiquated (and possibly useless as Lumsden suggests), but it emphasizes exactly the kind of group-oriented measurement that needs to be deemphasized if we can ever hope to make tests more relevant to individuals of varied backgrounds.

In essence, all bias in testing is caused by the failure of tests to take into account a number of important individual differences variables in their construction, administration, and scoring (Sattler, 1974; Angoff, 1975; Green, 1976; Pine & Weiss, 1976). These variables include individual differences in motivation, ethnic background and related variables.

Tests based on the classical model, because they are based on group statistics and because they require all testees to take identical test items, ignore individual differences. Therefore, trying to resolve the problems of test bias with the classical model is like trying to put a satellite into orbit with a submarine. What is needed if we are to make any progress in this critical research area is a theory that permits the testing process to be tailored to individuals. This technology now exists in the form of item response theory.

This research is supported by Contract N00014-76-C-0244, NR No. 150-383, with the Personnel and Training Research Programs, Office of Naval Research.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
Steven Pine
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) AND USERS OF THE ERIC SYSTEM."

This paper will discuss some of the research I am conducting at the University of Minnesota, supported by the Personnel and Training Programs of the Office of Naval Research (NR 150-383), to apply item response theory to the problem of test bias. I will begin by defining bias in terms of item response theory and indicate how the definition leads to several testable implications. Next, I will demonstrate these concepts with data from a real set of test items. Just to whet your appetite, I will tell you now that these items turn out to be unbiased! Finally, I will briefly discuss some of the current research we are conducting on incorporating some of these ideas with computerized adaptive testing to examine their potential benefits in testing minorities.

An Item Response Model of Bias

Item response theory. Recently, a new test theory called "item response (or latent trait) theory", which was specifically designed for the measurement of individuals, has emerged. Briefly stated, item response theory is based on the idea that the response which an individual makes to a given test item is determined by his/her ability on one or more underlying dimensions (latent traits), and the parameters of the test items, i.e., their difficulty, discriminating power, and probability of being guessed correctly by chance. This idea is expressed mathematically by the Item Characteristic Curve (ICC) which gives the probability that a testee of a given ability level on the underlying dimension will answer the administered item correctly.

The ICC curves and their associated item parameters are the building blocks of item response theory. Once item parameters are determined for each test item, they completely describe how individuals at a given ability level will perform on each item. This allows probabilistic statements to be made about the ability of a testee regardless of subgroup membership or which subset of items has been administered. This property provides a means for creating tests which can be adapted to individual testees, since with this test model one is no longer required to administer identical items to every testee. It is this property that makes item response theory so potentially valuable for developing less biased tests. Furthermore, the bias-reducing potential of item response theory is not tied to its use with any particular testing strategy, although the greatest benefits can be expected when it is used in conjunction with adaptive testing.

I have claimed that item response theory can be useful in constructing less biased tests. I will now define exactly what I mean by "unbiased", state the assumptions that must be met, and indicate some testable implications of these assumptions.

Definition of item bias. A test item is unbiased if all individuals having the same underlying ability have an equal probability of getting the item correct, regardless of subgroup membership.

As has been previously mentioned, the ICC gives the probability of getting an item correct given an ability level. Therefore, the above definition of an unbiased item is equivalent to requiring that a test item have the same ICC for all subgroups. Since an ICC is determined by its parameters, this is also equivalent to requiring that the item parameters which describe an item's difficulty, discrimination, and probability of being guessed correctly by chance be invariant up to a linear transformation from subgroup to subgroup. (Linear transformation is necessary to account for the fact that subgroups in which the parameters are calibrated may have ability distributions with different means and variances.)

Assumptions and tests for unbiased items. In this paper, I will restrict my discussion to tests that consist entirely of homogeneous items. By homogeneous, I mean items which measure approximately one ability dimension. The reason I say approximately is that I am allowing for the possibility that these items may measure one or more extraneous dimensions as well. For instance, test items intended to measure vocabulary ability may inadvertently also measure several cultural variables. By restricting ourselves to homogeneous items, we will have an easier time developing a theoretical foundation that can later be extended to the multidimensional case.

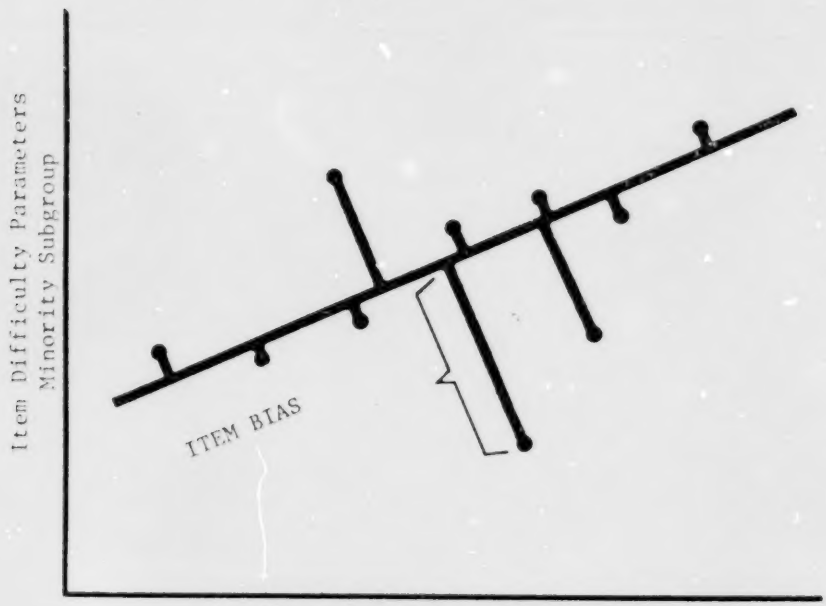
Lord & Novick (1968, p. 359) have shown that the parameters of a test item will be invariant across subgroups if the same dimension governs performance on that item for each of these subgroups. According to the definition given earlier, this implies that test items are unbiased if they measure the same single trait for all subgroups. This fact suggests that the bias in a test can be directly assessed by factor analyzing the matrix of inter-item correlation coefficients for each subgroup in the population and demonstrating that the non-random variance in each matrix is accounted for by a single factor.

Furthermore, if the same dimension underlies a set of test items for a population of testees (which would, therefore, make the items unbiased), the item parameters for any two subgroups in the population ought to have a linear relationship (Lord & Novick, 1968, p. 380). This condition can be tested directly by plotting the discrimination (a), difficulty (b), and guessing (c) parameters of one subgroup against another and testing for linearity. A plot of this type is shown in Figure 1.

Also, by comparing the relative item parameter values between subgroups, the specific test items which contribute the most to causing the item by group interaction to be nonlinear can be identified. In Figure 1, the perpendicular distance between each item and the best fitting line through all the points can be interpreted as the degree of item bias. Similar plots and interpretations could also be made for the item discrimination and guessing parameters.

The degree-of-item-bias index could be used to screen out the most biased items during the construction of a conventional test, or could be used within an adaptive testing framework as an additional item selection criterion.

Figure 1
Item Bias Shown as a Perpendicular Distance
in a Scatter Plot of Subgroup Item Difficulties



The procedure of assessing item bias by plotting a scatter diagram of item parameters for one subgroup against another is not in itself new. A very similar method has been used at Educational Testing Service (ETS) for years. The essential difference between the present method and the ETS method is that the ETS method uses item parameters based on classical test theory. It can be shown (Lord & Novick, 1968, p. 381) that classical item parameters will not, in general, be linearly related across subgroups of a population. This means that the test for bias using classical parameters can lead to an artifactual detection of bias. Furthermore, the difficulty parameter of classical test theory is confounded by level of discrimination and guessing effects (Urry, 1974). Therefore, if an item is relatively more difficult for one subgroup than another, one really does not know whether this is because the item varies on difficulty, discrimination, guessing, or some combination thereof.

The item response parameters, on the other hand, provide relatively unconfounded measures of difficulty, discrimination, and guessing. Therefore, by plotting these parameters on separate plots, one can determine exactly why an item is biased. For instance, for any given item it may turn out that an item is biased not because it is relatively more difficult for a minority subgroup, but because this subgroup is less effective at guessing. This kind of detailed analysis is impossible using classical item parameters.

An example with real data. To demonstrate how these analyses might be used and interpreted, I have applied them to 75 multiple-choice vocabulary items collected at a racially mixed high school in Minneapolis. The sample sizes in this study were smaller (58 Blacks, 168 Whites) than we would normally like, but the data provide a good example of the technique.

First, we can test the unidimensionality assumption by factor analyzing the inter-item correlation matrices. A subset of 45 items was chosen and two tetrachoric correlation matrices were calculated on these items, one for the Black and one for the White subsamples. The matrices were then factor analyzed using the principal axis method. Eight factors were extracted from each matrix. The eigenvalues from the two factor analyses are shown in Table 1.

Table 1
Eigenvalues from Factor Analysis of Black and White Subgroup Item-Intercorrelation Matrices

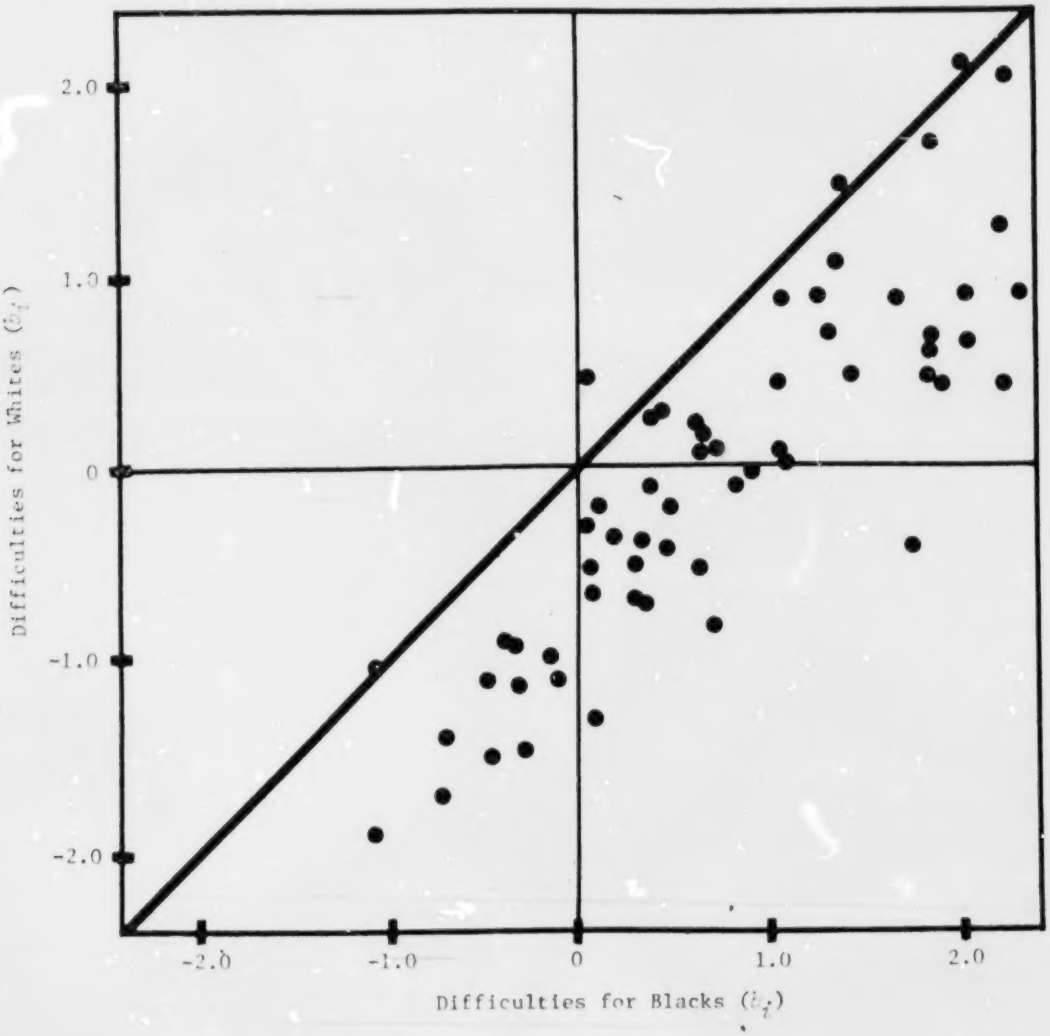
Subgroup	Factor	Eigenvalue	Percent of Variance	Cumulative Percent
Whites				
	1	19.26	64.8	64.8
	2	2.32	7.8	72.7
	3	1.67	5.6	78.3
	4	1.58	5.3	83.7
	5	1.37	4.6	88.3
	6	1.20	4.1	92.4
	7	1.18	4.0	96.4
	8	1.08	3.6	100.0
Blacks				
	1	16.33	47.9	47.9
	2	3.70	10.9	58.7
	3	3.01	8.8	67.5
	4	2.64	7.7	75.3
	5	2.35	6.9	82.2
	6	2.26	6.6	88.8
	7	2.06	6.0	94.9
	8	1.75	5.1	100.0

For both the Black and the White data, the first eigenvalue was very large and, therefore, the unidimensionality assumption seems to hold. Furthermore, the items appear to be measuring the

same dimension in both subgroups, because the coefficient of congruence (Rummel, 1970, p. 461) calculated between the 45 corresponding loadings for Factor I in the two subgroups was .97.

Now let's look at the results of the bias analysis carried out on all 75 items. The scatter plot in Figure 2 shows the result of plotting each item by its estimated difficulty parameter in the White subsample versus its estimated value in the Black subsample.

Figure 2
Race Contrasts of Item Difficulties for 75 Vocabulary Items



One of the first things to notice is that almost all of the items are relatively more difficult for Blacks than for Whites. This is indicated by the fact that the items tend to fall below the 45° angle line. The points would fall on this line if they were equally difficult for Blacks and Whites. However, in this analysis, the mere fact that the items are relatively more difficult for Blacks cannot necessarily be taken as an indication of bias. Bias in the test items is assessed by evaluating the degree of linearity in the plot.

For these items the relationship appears to be quite linear and, in fact, the correlation coefficient between Blacks and Whites is ρ=.86. This is consistent with the results of the factor analysis and indicates that these vocabulary items are essentially unbiased.

You may be curious to know what an unbiased set of test items looks like. I have, therefore, appended a sample of six of these items to this paper for your perusal. You will notice that there is nothing particularly special about any of these items. The items have not been specifically chosen to reflect minority cultures nor do they employ fancy response options. That is not to say that these methods are necessarily undesirable; the point, however, is that the extent to which items are biased can only be determined by an empirical investigation of how the items behave in each of the subgroups of the population.

If a more in-depth analysis of these items were desired, one could produce and interpret scatter plots for the item discrimination and guessing parameters in a similar manner. This will not be done in this case because reliable estimates of these parameters cannot be obtained with the sample sizes that were available. Another analysis that could be performed on the plot in Figure 1 is to determine the index of bias for each item. It is possible that even though the items taken as a group are unbiased, one or more of the items taken individually might be biased. For instance, in this data set, several items appear to have unusually large departures from linearity. Of course, it is possible that these large departures may be due only to sampling error. To eliminate possible misinterpretations that would occur if this were the case, I am working on a technique to establish confidence limits for the best fitting line so that we can say with some known degree of confidence that any item which exceeds these limits is biased.

Concluding Remarks

The material presented here is only a small example of how item response theory can potentially be applied to the problem of test bias. In fact, it represents only a small part of the research currently underway at the University of Minnesota. Nonetheless, I believe that this material provides a simple, but cogent example of the efficacy of applying item response theory to the bias problem.

In the time which remains, I would like to briefly describe some of our other research projects. The first project is really an extension of the research described here. It involves a method of correcting for bias in the item parameters. Very briefly, this method consists of determining item parameter estimates so that they will depend only on the extent to which an item loads on the factor it is supposed to be measuring.

In essence, this approach is based on the notion that to obtain unbiased test items, it is not necessary to expend huge amounts of time and money searching for so-called "culturally-relevant" test items. It does not really matter what the test items look like. All that is necessary is to know how each test item behaves (i.e., what its parameters are) in the various subgroups which comprise our test population. If we know that an item is biased, we can correct its parameter values, e.g., with the method I have suggested, to account for this bias. Then, if we base our test scores not on the total number of correct answers, but on some function of the corrected parameter values, we can produce unbiased ability estimates.

We are now in the process of checking out this method for correcting item bias with computer simulations. In this way, the estimated corrected-item-parameter values can be directly compared to the known, true-item-parameter values. If the results of the study are favorable, we will have devised a means of eliminating the effect of bias in test items.

Does this mean that we can now write the final chapter on test unfairness? No, not at all. First, I am not so naive as to think that a solution to the bias problem will ever be universally accepted as long as differences exist in the mean test scores of various subgroups. Secondly, bias in the estimation of item parameters is only one source of possible unfairness in the test process. A test can be unfair for a myriad of other reasons, including those attributable to elements in the testing environment, and to the procedures used to select and administer test items.

To explore the possible psychometric influences on unfairness, we have been conducting a series of computer simulations to investigate how the item characteristics of a test interact with the choice of a testing strategy, to influence test fairness in selection.

Finally, I would like to mention the research we are currently doing in the Minneapolis Public School System to show that I test people as well as electrons. In this study, we are investigating just how well some of the procedures mentioned here hold up in a real test administration. We will also be using the study as a proving ground for our recently developed bias-reducing computerized adaptive test, and to replicate a previous finding that computerized tests increase the test-taking motivation of minority testees. Perhaps next year a report of the results of these research efforts will be possible.

Appendix: A Sample of Six Unbiased Test Items

rudimentary	1	impolite
	2	uncivilized
	3	elementary
	4	protected
	5	destructive
ecstatic	1	dotted
	2	fawning
	3	proper
	4	rapt
	5	sullen
abacus	1	computing device
	2	flower
	3	herb
	4	musical instrument
	5	carpentry tool
gnome	1	dwarf
	2	lizard
	3	ghost
	4	rock
	5	miner
capricious	1	frugal
	2	callous
	3	medicinal
	4	tropical
	5	whimsical
exemplary	1	laudable
	2	inside
	3	quick
	4	lithe
	5	magic

REFERENCES

Angoff, W.H. The investigation of test bias in the absence of an outside criterion. Paper presented at the NIE Conference on Test Bias, December 1975.

Green, D.R. Reducing bias in achievement tests. Paper presented at the NIE Conference on Test Bias, San Francisco, April 1976.

Lord, F.M., & Novick, M.R. Statistical theories of mental test scores. Reading, MA: Addison-Wesley, 1968.

Lumsden, J. Test theory. In Rosenzweig, M.R., & Porter, L.W. (Eds.), Annual review of psychology. Palo Alto, CA: Annual Reviews, Inc., 1976.

Pine, S.M., & Weiss, D.J. Effects of item characteristics on test fairness. (Research Report 76-5). Psychometric Methods Program, Department of Psychology, University of Minnesota, Minneapolis, 1976.

Rummel, R.J. Applied factor analysis. Evanston, IL: Northwestern University Press, 1970.

Sattler, J.M. Assessment of children's intelligence. Philadelphia: W.B. Saunders Company, 1974.

Urry, V.W. Approximations to item parameters of mental test models and their uses. Educational and Psychological Measurement, 1974, 34, 253-269.